® ASSESSING UNDERSTANDING IN STATISTICS

LIPSON, Kay
Swinburne University of Technology, Lilydale
Australia

Using a framework for assessing dimensions of understanding in statistics a series of assessment tasks were developed by the researcher to address both procedural and conceptual understanding. This paper describes these tasks, together with the results of students' performance on the tasks. It will be shown that, while some of the tasks developed did assess the dimension of knowledge which they were developed to address, and some did not, overall it was possible to develop tasks to specifically assess both procedural and conceptual knowledge in statistical inference.

INTRODUCTION

Assessment in statistics has become of great interest to researchers in recent years, and considerable work has been done to develop a range of assessment instruments (Gal & Garfield, 1997). In particular, educators are interested in tasks which measure both procedural understanding, a students ability to correctly perform a task, and conceptual understanding, their knowledge of what they are doing and why they are doing it (Garfield, delMas, & Chance, 2002). This paper describes a variety of tasks which were developed to measure conceptual understanding, and provides some empirical evidence that while several of the tasks do indeed measure aspects of a student's conceptual understanding as intended, some do not.

FRAMEWORK FOR MEASURING UNDERSTANDING

Several researchers have developed theoretical models for thinking about the development of understanding in mathematics and statistics. Most recent is the hierarchy of statistical literacy, statistical reasoning, and statistical thinking described by Ben-Svi and Garfield (2004). The theoretical framework used in this paper is based on some earlier work by Putnam, Lampert and Peterson (1990), who described five dimensions of understanding: representation, knowledge structure, connections between types of knowledge, active construction of knowledge and situated cognition. Tasks that fall within the classification of understanding as representation are considered here to measure procedural understanding. Tasks which fall within any of the other four dimensions of understanding are considered to contribute to the measurement of conceptual understanding.

Based on the application of this framework by Nitko and Lane (1990) to understanding in statistics, the following assessment framework for statistics is suggested:

Procedural Understanding	
Understanding as representation	Tasks which involve application of standard notation, representation and algorithms to solve statistical problems. This would include standard applications of the *t*-test or chi-square test for example.
Conceptual Understanding	
Understanding as knowledge structure	Tasks which give insight into the knowledge structures of students. That is, tasks that demonstrate that the student has made a connections between concepts, such as hypothesis-testing and confidence intervals for example.
Understanding as connections between types of knowledge	Tasks that require students to integrate formal knowledge with informal knowledge developed outside the class. This would include tasks requiring the interpretation of statistical concepts.

Understanding as the active construction of knowledge	Tasks that enable the teacher to monitor the development of knowledge over time, such as concept maps.
Understanding as situated cognition	Tasks which require the student to apply their knowledge in a variety of contexts, different from those previously seen and discussed in the classroom.

ASSESSMENT TASKS AND STUDENT PERFORMANCE

Using this framework the following tasks were developed to measure conceptual understanding in introductory inference. For each task a rationale is given, together with the results obtained by a group of 23 students who attempted each question.

Sampling: Adapted from the Statistical Reasoning Test (Konold & Garfield, 1993) this task is designed to investigate whether students accept sampling as a valid method of obtaining information about a population and realise the implications of sample size.

A survey is conducted with a random sample of 282 university students, in order to find out how far they travel to university each day. One student questions the validity of the study, noting that there are 4000 students at the university, not just 282. Read each of the statements listed below carefully, and select the ONE response that sounds the most reasonable to you.

A Agree, 282 is too small a percentage of the 4000 (7%) to allow us to draw conclusions.

B Agree, you should have a sample that is at least 50% of the population in order to make inferences.

C Agree, they should get all the students to participate in the survey.

D Disagree, 282 is a large enough number to use for these purposes if the sample was a random sample of students.

E Disagree, if the sample is random, the size doesn't matter.

The best alternative is D, showing that the student has conceptually linked samples and populations, and appreciated that size of the sample is not related to the size of the population. A and B suggest that the student erroneously believed that population size was important, C rejects sampling as a valid method of investigating a population altogether, whilst E suggests that sample size is not important. This task is classified as *Understanding as knowledge structure*. Student responses are as follows.

Table 1
Student responses to the Sampling task

Alternative	Frequency
A	1
B	0
C	0
D	21
E	1

The notion that sampling is a valid method of obtaining information about a population appears reasonable for most students.

Hospital: This task concerns the relationship between sampling variability and sample size, and was developed by Tversky and Kahneman (1982).

Half of all newborn babies are girls and half are boys. Hospital A records an average of 50 births a day. Hospital B records an average of 10 births a day. On a particular day, which hospital is more likely to record 80% or more female births?

A Hospital A (with 50 births a day)
B Hospital B (with 10 births a day)
C The two hospitals are equally likely to record such an event.

Tversky and Kahneman found that 56% of undergraduate students incorrectly gave the answer C, suggesting that many believe the variability of the sampling distribution is independent of the sample size. Selecting B indicated that the student appreciates that the variability in the sampling distribution is larger when the sample size is smaller. Response A implies that the variability of the sampling distribution increases with the sample size, and may indicate confusion between the sampling distribution and the distribution of the sample.

This task requires the integration of formal knowledge with informal knowledge developed outside the class, an example of *Understanding as connections between types of knowledge* and contributing to the measurement of conceptual understanding. Student responses to this task are shown below.

Table 2
Student responses to the Hospital task

Score	Frequency	Percentage
A	1	4.3
B	6	26.1
C	16	69.6

Most students (69.6%) selected response C, slightly higher that the 56% observed by Tversky & Kahneman (1982). These students may not have explicitly linked the variability of the sampling distribution and the size of the sample in their conceptual structure for sampling distribution, or alternatively, the scenario used to illustrate sampling here does not evoke that conceptual link.

Confidence interval: This task is designed to establish whether or not students see hypothesis-testing and confidence intervals as alternate ways of looking at the same problem. In this instance that is achieved by determining whether or not the hypothesised value for the mean lies within the given confidence interval, and thus identifying any inconsistency in the two sets of conclusions.

Using a computer package, the student finds the 95% confidence interval for the mean number of residents to be (1.626, 3.465). Is this confidence interval consistent with the conclusion to the hypothesis test carried out in part (a)* Explain.

Part (a) refers to a *t*-test.

Ability to explain the relationship between the conclusions based on the hypothesis test and the confidence interval indicates the student has conceptually linked these two aspects of statistical inference, indicating *Understanding as knowledge structure*.

The results obtained showed a poor general recognition of the relationship between these two facets of statistical inference, with only eight students (35%) appreciating that the results were consistent because the confidence interval given did not include the hypothesised value of μ. The student who obtained 1 mark indicated that there was a contradiction, but then gave the correct reasoning. Of the students 14 who scored zero, three students either misinterpreted or omitted the question, whilst an alarming 11 students indicated that there was no inconsistency, as the sample mean of 2.545 lay within the confidence interval given, evidence of a lack of conceptual understanding of the confidence interval.

The scores achieved by the group of students are summarised below.

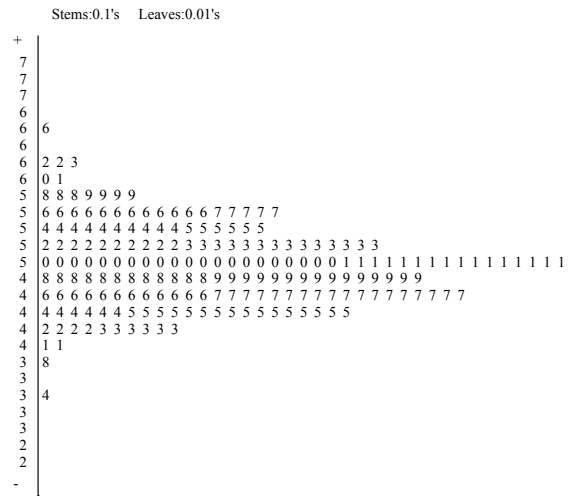
Table 3
Student responses to the Confidence Interval task

Score	Frequency	Percentage
0	14	60.9
1	1	4.3
2	8	34.8

Modelling: This task was designed to determine whether students were able to solve the same hypothesis-testing problem using both an empirical sampling distribution, generated by repeated sampling, and the normal model for the theoretical sampling distribution.

A sample of 100 primary school children were asked which type of protection they preferred to use to protect their faces from the sun, a hat or sunscreen. Of the 100 children, 61 preferred to use a hat, and 39 preferred to use sunscreen.

(a) Use the computer generated sampling distribution given to test if there is a difference in the proportion of students who show preference for a hat.



Stemplot showing 200 values of the sample proportion obtained from drawing samples of size 100 from a population with proportion $p = 0.5$.

(b) Use the normal model for the sampling distribution to test if there is a difference in the proportion of students who show preference for a hat.

Ability to carry out the hypothesis test using the empirical sampling distribution is taken as evidence of a conceptual link between the hypothesis-testing procedure and the sampling process. This aspect of the task is concerned with conceptual understanding in the category *Understanding as knowledge structure*. Carrying out the hypothesis test using the normal model, however, is a procedural task involving the use of formulae and tables, disassociated with the sampling process and evidencing *Understanding as representation*, which is procedural.

However, recognizing the equivalence of the two parts of the task, and in particular the expectation of consistency between the results of the two analyses falls into the category of *Understanding as connections between types of* knowledge, again as aspect of conceptual understanding. It was possible that some students would not be able to complete one part of the task but successfully complete the other, as they are measuring different aspects of understanding. Each part of the tasks was scored out of maximum of 7.

An analysis of the marks obtained by the students revealed that the mean score for part (a) was 5.0 marks whilst the mean mark for part (b) was 5.7 marks. A paired *t*-test showed that the students performed significantly better in part (b) than in part (a), ($t(22) = 2.296$, $p = 0.032$). This difference is not large, but indicates that the students have on the whole performed better in the task that measure procedural understanding than that measuring conceptual understanding. However, most students scored quite well on both parts of the task, with 17 students (74%) scoring 5 or more for part (a), and 20 students (87%) scoring 5 or more for part (b).

The correlation between the scores was also quite high ($r = 0.724$) indicating that, in general, those students who attained the higher scores for part (a) also attained the higher scores for part (b).

Unknown test: This task required the student to recognise a novel scenario as a hypothesis-testing situation, and identify the key components of the problem. The task was concerned with statistical testing as applied to the variance. Whilst the students were familiar with the notation for population variance (σ^2), they had not carried out any testing concerned with variance, nor seen an F-statistic before, in this course.

A researcher wishes to know whether blood pressures became more variable after a particular treatment. To determine this she carries out an F-test (which you have not been taught) which can be used to test for the equality of variance in two independent samples.

For the treatment group (n =10) the sample variance was found to be $s^2 = 76.44$ while for the placebo group (n = 11) the variance was $s^2 = 34.82$ giving an F statistics of 2.20, and a P-value of 0.26.

Write down appropriate null and alternative hypotheses and use the P-value to draw a conclusion about the variability of blood pressure in the two groups.

Successfully identifying this problem and relating the key features of the new situation correctly to those previously studied provide evidence the existence of a generalised schema for hypothesis-testing. Students who were unable to successfully complete this task may have constructed separate schemas for various examples of hypothesis-testing, but not successfully synthesised these schemas into a generalised schema for hypothesis-testing where each individual testing scenario is seen as an example of the over-riding principle. Such integrative reconciliation reflects conceptual understanding of this aspect of statistical inference. The task thus can be considered as measuring *Understanding as knowledge structure*.

Only five students were able to present a correct solution to this problem, and of these 5, only two actually wrote their hypotheses in symbols, whilst the other three wrote them in words. This would perhaps indicate some uncertainty about which symbol to use, as almost all hypotheses in this teaching sequence are written in symbols. Of the rest, 11 wrote their hypotheses incorrectly in terms of *s*, and another four wrote their hypotheses in terms of μ. One student wrote hypotheses in terms of F, another ρ and another felt unable to complete the hypotheses in this question. Regardless of the hypotheses, all but one student used the P-value correctly to make a decision regarding the appropriateness of the null hypothesis in this context.

The scores given in this question are shown in below. The maximum available score was four marks.

Table 4
Student responses to the Unknown task

Score	Frequency	Percentage
1	3	13.0
2	15	65.2
3	0	0
4	5	21.7

Explanation: This task was required students to interpret the (procedural) steps in the hypothesis test (given to them as shown below) in their own language, as if explaining to a person with no previous training in statistics.

Steps in your hypothesis test	Explanation
Hypotheses: H_0: $\rho = 0$ H_1: $\rho \neq 0$ non-directional test *Significance level:* $\alpha = 0.05$ *Test statistic* $r = 0.5135$ for $n = 30$ pairs of data values *P-value* P-value $= 2 \times P(r > 0.5135)$ $\approx 2 \times 0.0025$ $\approx$ 0.005 or 0.5% *Decision & conclusion* As $P < 0.05$, reject H_0 and conclude that there is a relationship between the intelligence of children and their mothers in the general population.	

By linking the formal notation and algorithms with informal knowledge which can be understood by most people students are demonstrating *Understanding as connections between types of knowledge.*

Marks were not awarded where the student merely re-phrased or described in words the step in the hypothesis test. The task was scored out of a maximum of 10, and the distribution of results achieved by the group was symmetric, ranging from 1 to 9. The mean score was 4.7 and the standard deviation 2.2.

Analyses of the responses indicated that most students knew what they were trying to achieve by carrying out the hypothesis test, and were correctly able to interpret the conclusion. However, the general understanding of level of significance and P-value was very much lower. Many students wrote a procedural definition of the level of significance, such as "we will reject the null hypothesis of the P-value is less than this value", whilst operational interpretations of the P-value such as "to get the P-value you look up the tables and then multiply the answer by two because it is a two tail test" were quite common.

Radio: This task is designed to ascertain which aspects of their statistical knowledge students are able to relate to this real world context. The question is quite intentionally open ended, and contains insufficient information for an exact answer to be obtained using a standard algorithm.

A radio station claims to its advertisers that 20% of 18–25 year olds listen to this station between 6.00 pm and mid-night on weeknights. A market research company carries out independent research on behalf of an advertiser and finds that only 15% of their sample of 18–25 year olds listen to the radio station in this time period. The advertiser concludes that the radio station is misleading them. What do you think? Try to include all the relevant reasons for your answer.

A complete discussion would include both issues concerned with the sampling process, and issues concerned with sample size. Consideration of the sampling process alone, explanations implying that there must be a problem with the data collection, suggest that the student has not recognised the link between sampling and the sampling distribution in this scenario. Consideration of the sampling distribution suggests that the student recognises the plausibility of the difference between the population parameter and the sample statistic. Further identifying the important role of sample size suggests that the student's conceptual structure for sampling distribution includes recognition of the role of sample size in explaining sampling variability. Overall this task can be classified as measuring *Understanding as situated cognition*. The general themes of the students' answers are summarised below.

Table 5
Student responses to the Radio task

Theme of answer	Frequency
Sample selection problem only	6
Hypothesis test only	8
Both	9

The average mark achieved in this question was 3.1 out of a maximum of 5. Of the 23 students in the group, a total of 17 or 73.9% recognised that sampling variability was a possible explanation for the difference between the sample statistic and the population parameter and that a statistical procedure existed which would enable them to decide if the this was the preferred explanation. This indicates that for these students, the hypothesis-testing procedure was well enough understood to be recognised in a real world situation.

DO THE TASKS DEVELOP USING THE FRAMEWORK MEASURE CONCEPTUAL UNDERSTANDING?

These tasks were designed to measure conceptual understanding in statistical inference. The intention was to combine the student scores on these and another set of tasks measuring procedural understanding to assign each student two separate scores, one measuring procedural understanding and one measuring conceptual understanding. To achieve these composite variables, factor analysis (Klein, 1994) can be used to reduce a large number of inter-related variables to a relatively small number of underlying factors which are conceptually meaningful.

An exploratory factor analysis was carried using scores for all of these tasks together with six standard tasks designed to measure procedural understanding. All of the tasks designed to measure procedural understanding loaded onto the same factor, and most of the tasks designed to measure conceptual understanding loaded onto a second factor. However, there were some unexpected results. Both parts of the Modelling question were highly correlated with other questions measuring procedural understanding. On reflection this result was understandable, as questions similar to the Modelling task had been regularly discussed during the course, and given as practice exercises out of class meaning that students had a considerable amount of previous practice with similar tasks.

The Sampling and Unknown tasks also loaded onto the factor measuring Procedural understanding, whilst the Confidence Interval task was not loading highly onto either factor. Detailed analysis of the students' responses to the Unknown task revealed that for this group of students, the scenario presented was not novel for all students. For some who had undertaken previous courses in statistics, this task was already known and was thus be classified as

procedural understanding. For others, it fell into the domain of conceptual understanding, as they had not seen this before. On this basis, the decision was made to repeat the factor analysis without the Unknown task. Similarly, it was concluded that on the basis of the student's previous statistical experience, the Sampling task would be procedural for some students and conceptual for others, and thus this task was omitted from the final analysis.

After these tasks had been removed the data was found again to be suitable for factor analysis, with KMO = 0.801, and Bartlett's Test of Sphericity returning a P-value less than 0.0005. The resultant factors attained through the second analysis together explained 67.9% of the total available variance. The factor matrix showed simple structure, and the resultant factors were clearly identifiable as measures of procedural understanding (Factor 1) and conceptual understanding (Factor 2). The correlation between Factor 1 and Factor 2 was 0.254, indicating a weak positive relationship between the two factors, due perhaps to a general underlying ability factor.

CONCLUSION

These analyses confirm that, as suggested by the theoretical analysis, the tasks developed concerning assessment of aspects of understanding can be resolved into two variables measuring the underlying constructs of procedural and conceptual understanding. They also indicate, however, that the theoretical analysis alone is not sufficient to establish the dimension the question is addressing. However, the establishment of two largely independent factors indicates that educators should ensure that they are assessing both dimension of statistical understanding.

REFERENCES

Ben-Zvi, D., & Garfield, J. (2004). Statistical Literacy, Reasoning, and Thinking: Goals, Definitions and Challenges. In D. Ben-Svi & J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning, and Thinking* (pp. 3-15). Dordrecht: Kluwer Academic Publishers.

Gal, I., & Garfield, J. (1997). Curricular Goals and Assessment Challenges in Statistics Education. In I. Gal & J. B. Garfield (Eds.), *The Assessment Challenge in Statistics Education* (pp. 1-16). Amsterdam: IOS Press.

Garfield, J., delMas, R., & Chance, B. (2002). *The Assessment Resource Tools for Improving Statistical Thinking (ARTIST) Project*. Retrieved 7 January, 2007, from https://app.gen.umn.edu/artist/

Klein, P. (1994). *An easy guide to factor analysis*. London: Routledge.

Konold, C., & Garfield, J. (1993). Statistical Reasoning Assessment. Part 1: Intuitive Thinking, *SRRI, University of Massachusetts, Amherst, MA*: Unpublished manuscript.

Nitko, A. J., & Lane, S. (1990). Solving Problems is Not Enough: Assessing and Diagnosing the Ways in which Students Organise Statistical Concepts. In D. Vere-Jones (Ed.), *Proceedings of the 3rd International Conference on Teaching Statistics* (Vol. 1, pp. 467-474). Dunedin, NZ: International Statistics Institute.

Putnam, R. T., Lampert, M., & Peterson, P. L. (1990). Alternative Perspectives on Knowing Mathematics in Elementary Schools. In C. B. Cazden (Ed.), *Review of Research in Education* (pp. 57-150). Washington, DC: AERA.

Tversky, A., & Kahneman, D. (1982). Judgement under uncertainty: Heuristics and biases. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgement under uncertainty: Heuristics and biases* (pp. 3-20). Cambridge: Cambridge University Press.